

A Framework for Reproducible, Interactive Research: Application to health and social sciences

Joao Ricardo Nickenig Vissoci¹, Clarissa G. Rodrigues^{2,3}, Luciano de Andrade⁴,
Jose Eduardo Santana⁵, Amrapali Zaveri⁶, and Ricardo Pietrobon²

¹Inga College - Medicine Department

²Duke University Medical Center, {jnv4, clarisa.rodrigues, rpietro}@duke.edu

³Instituto de Cardiologia do RS, Fundacao Universitaria de Cardiologia, Rio Grande do Sul, Brazil

⁴Universidade Estadual do Oeste, do Parana, Brazil,
luciano.andrader@unioeste.br

⁵Instituto de Computação, Universidade Federal de Alagoas (UFAL), Alagoas, Brazil, jes@ic.ufal.br

⁶Universität Leipzig, Institut für Informatik, D-04103, Leipzig, Germany,
zaveri@informatik.uni-leipzig.de

April 23, 2013

Abstract

The aim of this article is to introduce a reporting framework for reproducible, interactive research applied to Big Clinical Data, based on open source technologies. The framework is constituted by the following three axes: (i) data, (ii) analytical codes and (iii) dissemination. In this paper, different documentation formats and online repositories are introduced. To integrate and manage the reproducible contents, we propose the R Language as the tool of choice. All the information is then published and gathered in a website for different projects. This framework is free and user friendly and is proposed to enhance reproducibility of health-science reports.

1. INTRODUCTION

With the growing amount of data in healthcare, the ability to analyze large datasets and report results adequately has become a key factor of research and innovation [1], which supports the creation of new technologies and improved clinical decision making. The increased complexity of these datasets brings together difficulties and new challenges in terms of data management,

modeling and communication. Therefore, investigators are now focusing on developing reproducible research protocols including entirely reproducible data analysis. It implies that the results reported in a publication can be immediately reproduced by granting access to both the datasets as well as the statistical and data mining scripts of the study [2]

In order to make the information widely usable, the value of data collection, analysis and communication as well as the use of common

standards for sharing information have been recognized. In addition to increasing dissemination and better understanding of research findings, data sharing can also support confirmation or refutation of research by allowing replication and increased transparency of results [2, 3]

However, data sharing does bring some implementation challenges and possible risks. Potential invasion of participants privacy and breaking of patients confidentiality are primary concerns when making datasets public. Secondly, adequate data management, academic and commercial primacy, and intellectual property rights as well as journal copyrights are factors to be careful with while publishing data [3]. In this context, the use of an adequate framework becomes essential to allowing reproducible research without compromising such aspects, specially when analyzing and reporting results from large datasets.

Thus, the aim of this article is to introduce a simple reporting framework for *reproducible, interactive research* applied to health and social scienc. The framework is constituted by the following three axes: (i) data (Section 2.1), (ii) analytical codes (Section 2.3) and (iii) dissemination (Section 2.6). In this paper, different documentation formats and online repositories are introduced. To integrate and manage the reproducible contents, we propose the R Language as the tool of choice. All the information is then published and gathered in a website for different projects. This framework is free and user friendly and is proposed to enhance reproducibility of health-science reports.

2. REPRODUCIBLE RESEARCH FRAMEWORK

The framework proposed in this paper is based on the concept that an appropriate reproducible research report should allow one to totally reproduce the methods applied. Thus, we understand that besides making the analytical data, code and figures available, an adequate reproducible research framework should integrate tools and features in a way that others

could reach the same results and understand the process behind it.

Therefore, in order to achieve an adequate integration between data, codes and outcomes (figures, tables, numerical results and others) in our framework, we utilize the R Language [4] as the central tool. R has the ability of integrating and managing different data formats, codes and formats. In addition, it allows communication with several other analytical softwares such as SAS, Stata and SPSS [5, 6, 7].

2.1 Data formats

The first issue about making a research protocol reproducible is the data management process. There are several ways of storing data and many different data formats. In our perspective, some of them are better by allowing integration with data analysis softwares and online repositories as well as their ease of use. In the following sections, we demonstrate some of these formats we have been using and their integration with our reproducible Framework.

2.1.1 Reproducible Data

When making datasets publicly available, one must be concerned with the information that is going to be made public. In this context, the Health Insurance Portability and Accountability Act (HIPAA) developed a section on Protected Health Information (PHI), which means that individually identifiable health information must be kept confidential when sharing data in healthcare. The complete list of PHI can be found at the Health and Human Services US Department [8]

Secondly, it is important to make sure that the data is coded with appropriate names that allows other people to read and understand the content easily. To make it easier, we strongly encourage the publication of a complete and organized *data dictionary* together with the dataset, containing variable labels, respective code, data characteristics (continuous, discrete, ordinal, dichotomous, etc.) and any other source of relevant information (e.g. length of Likert scale, categorization factors).

2.1.2 CSV

Comma separated values (CSV) is a format readily available for consumption by any data analysis language or software. However, it does not provide a way to update the data once it is downloaded other than downloading the dataset again. In addition, the CSV format does not offer any security features.

On the other hand, CSV files have one of the best usability experiences among all the formats and it can be easily integrated with R using online repositories (i.e. Google Drive or Dryad) through different R packages. One such package is the RCurl package [27], which can integrate R with different HTML domains, among them a .csv spreadsheet from Google Docs.

2.1.3 RDE, LOD and SPARQL endpoints

Semantic Web technologies have recently become popular given the success provided by Linked Open Data (LOD) [10]. The data is represented with the help of the Resource Description Framework (RDF) format, while data sets themselves are queried through the SPARQL (a recursive acronym for SPARQL Protocol and RDF Query Language). Main advantages include the data availability 24/7 with automated updates and also the ability to dynamically merge across data sets sharing identical elements (classes or instances).

RDF data can be easily integrated with R Analytical codes through the RRDF package [11]. This package allows users to perform SPARQL queries inside R's workspace. In addition to this package mentioned there is a whole set of tutorials and packages that can be used within R [4].

2.1.4 JSON

JavaScript Object Notation (JSON) is considered one of the best data-interchange formats. It is a text format with conventions familiar to several programming languages such as C++, Java, JavaScript and Python.

More information and specifications about

how to integrate JSON data with specific applications can be found at [12]. Its connection with R analytical code is executed through the rjson package [13] which converts JSON objects into R objects.

2.2 Data repositories

After deciding a format of data to be used, it is also mandatory to use an online repository to store the data and integrate it with the analytical codes (discussed later in Section 2.3). In the following sections, we present some options of free repositories that are used by our group.

2.2.1 Dryad

Dryad [14] is an international repository specific for data related to scientific publications. It allows data to be deposited easily and readily provides the citation related to the respective publication. Dryad can be integrated with R, thus improving interoperability [15].

2.2.2 Figshare

Figshare [16] is an online repository, similar to Dryad, that allows researchers to choose a publication with the ability to be cited within the paper.

Additionally, Figshare supports not only data but also other types of research outputs such as figures, datasets, media files, papers, posters or even file sets with different types of documents. A major advantage of Figshare is its ability of easily sharing and discovering information about different research projects. We have used Figshare to publish datasets (in .csv formats) as well as figures. Examples can be found in [17, 18]. In addition, Figshare can also be integrated with R through some packages [19].

2.2.3 Google drive

Google Drive is another online repository which facilitates collaboration and sharing of files [20]. This application from Google integrates texts, spreadsheets, presentations and

other editors from Google (i.e. Google Docs, Google Sheets, Google Forms and others) and also allows the user to store forms, drawing, and different types of files in the cloud.

Google Drive is extensively used to share data, codes and other outputs among researchers in our group [21]. One example to connect data stored in Google Drive with R is the RCurl package . This package allows users to compose general HTTP requests and call URLs and other web formats, such as datasets in .csv format. Another way is to simply open the files stored in Google Drive (Spreadsheets or R-Scripts, for example) inside R, through RStudio [22].

In addition, we also use Google Drive as a way to integrate and facilitate collaborative writing and coding in R, since this approach has been found more user friendly to content researchers than other more sophisticated repositories.

2.3 Analytical scripts

Publishing analytical codes is an important step in a reproducible framework besides the connection between the codes and the data. Therefore, we demonstrate here the different software that can be used to generate, publish and manage the analytical codes.

2.3.1 R Language Statistical Software

As mentioned before, R [4] is the central tool of our reproducible research framework. As a definition, R is an open source software for statistical analysis and graphic creation. It has been developed by a vast community of collaborators from several countries and institutions.

Although R is not superior to other statistical softwares in every aspects (such as intuitive GUI interface, or pre-defined operations), it gathers qualities which makes it a better option to our framework than other statistical environments. One major advantage of R is its collaborative function in the development of packages. R has a huge library [4] (Comprehensive R Archive Network - CRAN) of packages for statistical analysis, graphic creation, data

mining and management, and integration with other softwares and programming languages.

This collaborative ability, besides making R a powerful analytical environment, makes it assume a position in our framework as a glue for other languages and technologies such as Python, Java, relational databases, RDF, C, C++, Weka, among many others. This way we can gather data and data storage tools, analytical coding and repositories for outputs, making a research project fully reproducible. In addition, R has being used by a large community, and has a lot of references to lean on.

In our group we opt to run R through RStudio [22]. This platform is also open source and is an integrated environment that helps to visualize the different R interfaces (workspace, graph, scripts and log). Other than that, it facilitates the management of multiple working directories through the definitions of projects.

2.3.2 Reproducible Scripts

As suggested by Hadley in his github repository [23], the idea is to create a code that can be recreated just by copying the codes we publish online. Therefore, each code must be connected to the dataset and contain all the information needed to be performed.

The elements of a reproducible script in R include the required packages, connection to the data, codes and codes descriptions. Each function in R is called upon a package where it is nested. So, for anyone else to be able to reproduce our codes, she must have all the packages installed.

Regarding the data, we have already discussed earlier the possible formats and ways to publish it. It is noteworthy that the data must be aligned with the codes. This means that all the variables must be named exactly with the names used in the codes. Also, every data management information must be inserted in the codes so that whoever is trying to reproduce it might reach the same results. Finally, each line must have a description of its purpose and use.

2.3.3 Github

Github [24] is an online repository built to facilitate the collaborative writing of computing codes. It not only allows the sharing of codes but facilitates collaboration through the copy (hereby called "*fork*") of project pages in a safer way, regarding the original code. Among all the qualities of using Github as a reproducible strategy in the analytical coding process, we highlight its strong connectivity with R. It allows not only the sharing and management of codes in websites, but also simulates R outputs with knitr [25].

There are several possibilities of using R integrated with Github. We have been using Github mainly to:

- publish analytical R-Scripts
- promote collaboration among our data analysts when creating or debugging data analysis
- generate automatic data reports for open design projects
- create templates for data analysis (hereby called data analysis toolbox) with explanation of the methods (using wiki pages) and description of codes and outcomes.

2.4 Dynamic research

In order to have a complete reproducible script and also to facilitate data dissemination and visualization, it is important to obtain automated and dynamic representation of tables, figures and reports. R allows the creation of analytical codes that generates automated reports, such as the knitr package [25], which translates the analysis into an HTML report (or other formats such a PDF). In summary, this package translates the code into a report mixing Latex and markdown languages. An example of its application can be found in our Github repository [26] for the, Glocal Open Design Collection project. In this specific project we used knitr associated with a R code to generate an automated report about data quality and associations.

Another way of using R to generate dynamic research is by developing interactive graphs. These are graphs that might be customized or modified by the user (research subject, patient or any other stakeholder) to get different slices of the dataset. R has several ways of generating interactive graphs. Here, we would like to introduce rggobi and Shiny [27, 28, 29]. However, there are options that can be found at the CRAN task view for dynamic graphs [30]

2.5 Licensing

Since all the documentation we are using is going to be made public we need to assure that its use is covered by a license. This will assure that any use other than that allowed by the license, is not performed by the users. This is fairly important due to the relevance of the information being made public. In our framework, we have used Creative Commons, which is a free copyright license framework [31].

Inserting a line regarding the licensing characteristics in each of the documents in a project is sufficient to specify the type of license. The licensing assures the need for approval from the copyright owner. Basically, we allow the user to share and adapt the specific parts of the project. The only restriction is that the user must attribute the documents to the original authors and must use it only for noncommercial purposes.

Examples of licenses are: This code is licensed under a Creative Commons Attribution - Noncommercial 3.0 Unported License. You are free: to Share - to copy, distribute and transmit the work, to Remix - to adapt the work, under the following conditions: Attribution - You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work). Noncommercial - You may not use this work for commercial purposes. With the understanding that: Waiver - Any of the above conditions can be waived if you get permission from the copyright holder. Public Domain - Where the work or any of its

elements is in the public domain under applicable law, that status is in no way affected by the license. Other Rights - In no way are any of the following rights affected by the license: Your fair dealing or fair use rights, or other applicable copyright exceptions and limitations; The author's moral rights; Rights other persons may have either in the work itself or in how the work is used, such as publicity or privacy rights. Notice - For any reuse or distribution, you must make clear to others the license terms of this work. The best way to do this is with a link to this web page. For more details see <http://creativecommons.org/licenses/by-nc/3.0/> [31]

2.6 Data Dissemination and Communication

Other than discussing methods and tools to make a research reproducible, we also believe that it is important to include a facilitation of the data communication and dissemination. This will not only allow users to access the research project but will also catalyze the reach and dissemination of the respective projects.

In order to disclose and gather all the material from our groups' research projects that was made public, we created websites (using Google Sites [32]) for each of the projects where we included links to data repositories, code repositories and inserted reports and graphs. Any web design tool can be used but our choice for Google Sites is based on its free access and user friendly interface. An example is the Observer Agreement website which integrates all the reproducible documentation for our researchers with observer agreement about orthopedic scales projects [33].

2.7 Overall workflow

Summaryzing the information discussed, we created a simple graphical demonstration of the framework's conception (Figure 1). As mentioned before, R Language software gets a highlighted position in the framework's model. So R is used to manage and coordinate the docu-

mentation. Data is stored in open access online repositories, in a R supported formats that will allow the connection between data and analytical code. The analytical codes are developed within R interface and stored in an open access online repository. Outputs generated by the codes are also stored in open sourced online repositories. All this information is licensed and integrated in a website for the research project.

3. DISCUSSION

In this study we aim to introduce a reporting framework for reproducible and interactive research, based on technologies and methods applied to some of the recent projects in our research group (RoR). Several tools were described to publish datasets and analytical codes, all centered and managed by the R Language Software.

The concept of our framework was initially based on some guidelines already published [34, 35]. Not many reports can be found in the literature on the use of reproducible research framework in healthcare. Some researchers do publish their datasets or codes, but generally they are published separately. In our framework we tried to approach different aspects of reproducibility, rather than just the data, that is connectivity [36], dissemination and licensing with a framework that is constituted of free and friendly technologies, facilitating *replication* and improving *transparency* of results.

Some of the tools we showcased here have been extensively discussed and used for the development of projects by many investigators. Github, for instance, has been extremely used by data analysts and programmers, as well as Dyrad and Figshare, given the increased amount of data being stored in clouds. However, this advances have not been observed as often in healthcare research, specifically when it comes to Big Clinical Data and replication of health researches protocols [3].

Although our proposed framework is still in progress and needs to be improved, we em-

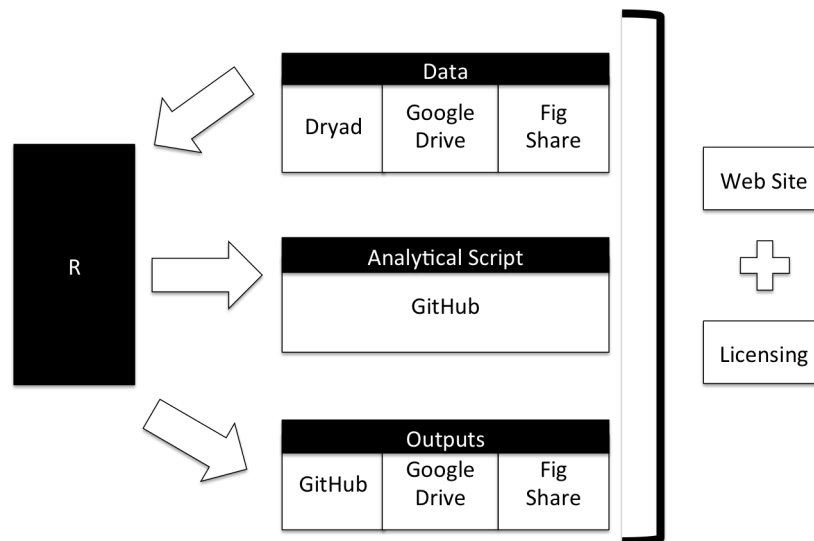


Figure 1: Depiction of the reproducible research framework.

phasize its ability not only for sharing data and codes in a safe way, but also connecting and disseminating information through free and user friendly technologies. We believe that only by sharing and comparing methods a consensus of framework can be created. Therefore, this model proposed can help towards the standardization of reproducible research protocols in healthcare, aggregating value not only for research, but also for innovation and clinical practice.

REFERENCES

- [1] Maniyka J, Chui M, Brown B, Bughun J, Dobbs R, Roxburgh C, Byers AH (2011). Big data: The next frontier for innovation competition, and productivity. McKinsey and Company. Available: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation. Accessed 18 April 2013.
- [2] Peng RD, Dominici F, Zeger SL (2006). Reproducible epidemiologic research. *Am J Epidemiol*. May 1;163(9):783-9.
- [3] Groves T, Godlee F (2012). Open science and reproducible research. *BMJ*. Jun 26;344:e4383. Available: doi:10.1136/bmj.e4383.. Accessed 18 April 2013.
- [4] R-project contributors. The R Project for Statistical Computing. Available: www.r-project.org, 2013. Accessed 18 April 2013.
- [5] IBM. SPSS Software .Available: <http://www-01.ibm.com/software/analytics/spss>. Accessed 18 April 2013.
- [6] SAS Institute Inc. Statistical Analysis System - SAS. Available: <http://www.sas.com>. Accessed 18 April 2013.
- [7] StataCorp LP. Stata: Data analysis and statistical software. Available: www.stata.com. Accessed 18 April 2013.
- [8] Health Information Privacy and Security (HIPAA). Available: <http://www.hhs.gov/ocr/privacy/hipaa>. Accessed 18 April 2013.
- [9] Lang DT (2013). Package 'RCurl'. Available: <http://cran.r-project.org/web/packages/RCurl/RCurl.pdf>. Accessed 18 April 2013.

- [10] Linked Data. Available: <http://linkeddata.org>. Accessed 18 April 2013.
- [11] Willighagen E (2013). Package 'rrdf'. Available: <http://cran.r-project.org/web/packages/rrdf/rrdf.pdf>. Accessed 18 April 2013.
- [12] JavaScript Object Notation (JSON). Available: www.json.org. Accessed 18 April 2013.
- [13] Couture-Beil A (2013). Package 'rjson'. Available: <http://cran.r-project.org/web/packages/rjson/rjson.pdf>. Accessed 18 April 2013.
- [14] Dryad Digital Repository. Available: datadryad.org. Accessed 18 April 2013.
- [15] Chamberlain S, Boettiger C, Ram K (2013). Package 'rdryad'. Available: <http://cran.r-project.org/web/packages/rdryad/rdryad.pdf>. Accessed 18 April 2013.
- [16] Figshare. Available: <http://figshare.com/>. Accessed 18 April 2013.
- [17] Moreira T, Yen T, Vissoci JRN, Barros T, Ejnisman L, Massa B, Pietrobon R, Vail TP (2013). Total Hip Arthroplasty Complications Prevalence Meta-analysis at 5, 15 and 20 years followup. Available: <http://dx.doi.org/10.6084/m9>. Accessed 18 April 2013.
- [18] Dal Ponte T, Pessin DV, Gambeta CE, Ferreira APB, Braga L, Vissoci JRN, Braga-Baiak A, Gandhi M, Pietrobon R. (2013) The reliability of AO classification on femur fractures among orthopedic residents. Available: <http://dx.doi.org/10.6084/m9>. Accessed 18 April 2013.
- [19] Boettinger C, Chamberlain S, Ram K, Hart E. (2012). Package 'rfigshare'. Available: <http://cran.r-project.org/web/packages/rfigshare/rfigshare.pdf>. Accessed 18 April 2013.
- [20] Google Drive. Available: <https://drive.google.com/>. Accessed 18 April 2013.
- [21] Research On Research and Innovation (ROR). Available: <https://sites.google.com/site/researchonresearchtech/home>. Accessed 18 April 2013.
- [22] RStudio Inc (2013). Available: www.rstudio.com. Accessed 18 April 2013.
- [23] Wickham H (2013). Reproducibility. Available: <https://github.com/hadley/devtools/wiki/Reproducibility>. Accessed 18 April 2013.
- [24] Github. Available: <https://github.com/>. Accessed 18 April 2013.
- [25] Xie Y. (2013). knitr: A general-purpose package for dynamic report in R. Available: <https://github.com/hadley/devtools/wiki/Reproducibility>. Accessed 18 April 2013.
- [26] Glocal Registry project in Github. Available: <https://github.com/rpietro/GlocalRegistry>. Accessed 18 April 2013.
- [27] Lang DT, Swayne D, Wickham H, Lawrence M. (2012). rggobi: Interface between R and GGobi. Available: <http://cran.r-project.org/web/packages/rggobi/index.html>. Accessed 18 April 2013.
- [28] Adler D, Murdoch, D. (2013). 3D Visualization package (OpenGL). Available: <http://cran.r-project.org/web/packages/rgl/index.html>. Accessed 18 April 2013.
- [29] RStudio Inc. (2013) Shiny: web applications framework for R. Available: <http://cran.r-project.org/web/packages/shiny/index.html>. Accessed 18 April 2013.

- [30] Lewin-Koh N. (2013) CRAN Task View: Graphic Displays and Dynamic Graphics and Graphic Devices and Visualization. Available: <http://cran.r-project.org/web/views/Graphics.html>. Accessed 18 April 2013.
- [31] Creative Commons. Available: <http://creativecommons.org/>. Accessed 18 April 2013.
- [32] Google Inc. Google Sites. Available: <https://sites.google.com/?pli=1>. Accessed 18 April 2013.
- [33] Observer Agreement. Available: <https://sites.google.com/site/observeragreement/home>. Accessed 18 April 2013.
- [34] Laine C, Goodman SN, Griswold ME, Sox HC. (2007) Reproducible research: moving toward research the public can really trust. *Ann Intern Med*. Mar 20;146(6):450-3.
- [35] Peng RD. (2009) Reproducible research and Biostatistics. *Oxford Journals. Biostatistics*. Volume10, Issue pp 405-408.
- [36] Peng RD. (2011) Reproducible Research in Computational Science. *Science* 2 December 334 (6060):1226-1227 DOI: 10.1126/science.1213847